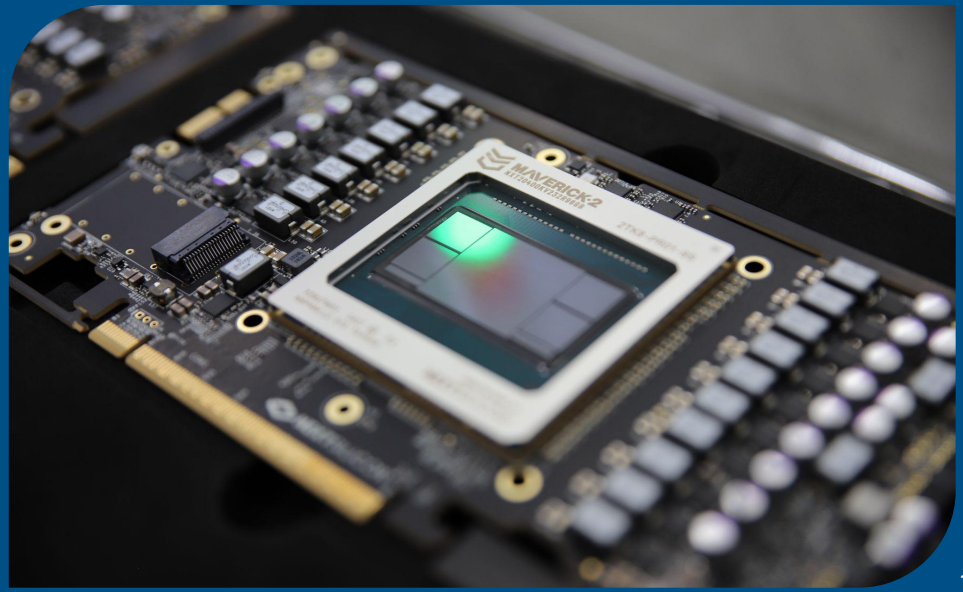
Next-Generation AI Silicon

Dr. Ian Cutress
More Than Moore



A New Chiplet Era

Legacy Hardware: CPU, GPU, FPGA

Old and New: Analog, Neuro, Optical

Reduced Precision

Silicon Market: Datacenter, Edge

AI Hardware: Startup Funding

AI Hardware: What's New

A New Chiplet Era

Legacy Hardware: CPU, GPU, FPGA

Old and New: Analog, Neuro, Optical

Reduced Precision

Silicon Market: Datacenter, Edge

AI Hardware: Startup Funding

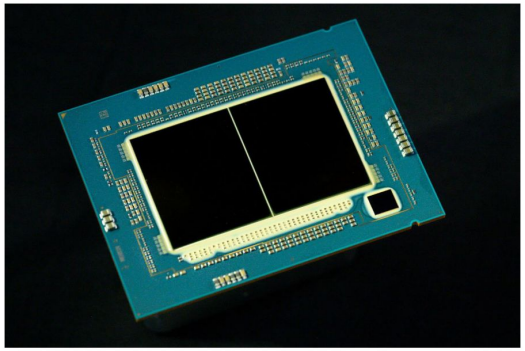
AI Hardware: What's New

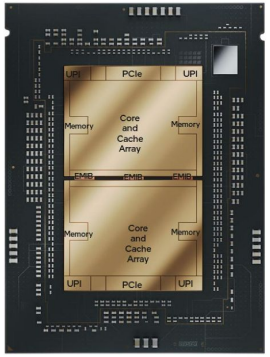
A New Chiplet Era

Legacy Hardware

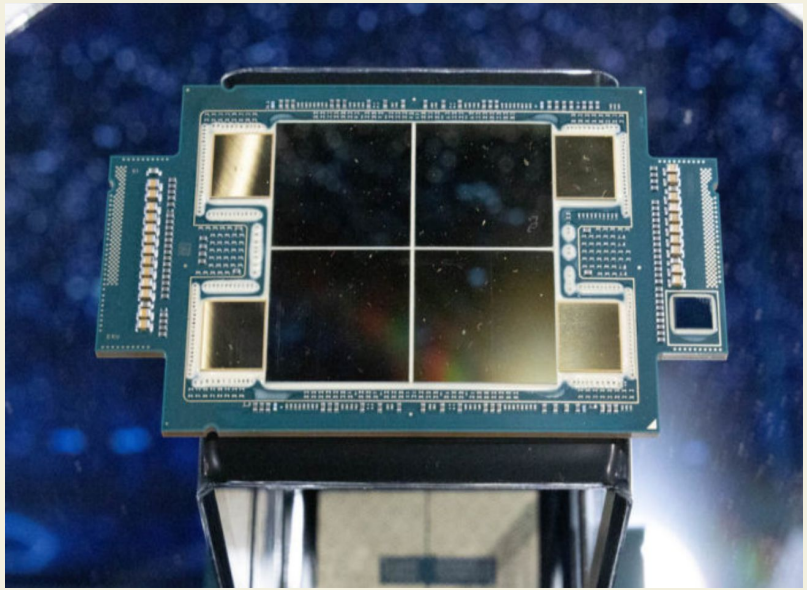
- CPU: x86, Arm, RISC-V
- GPU: NVIDIA, AMD, Intel
- FPGA: Intel (Altera), AMD (Xilinx), Lattice
- ASIC: SmartNIC, Anton3

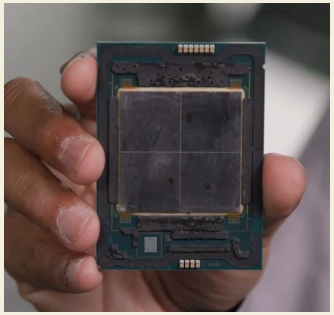
Intel CPU





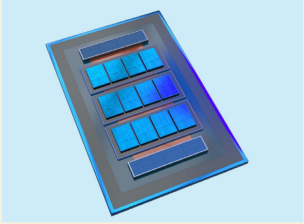
Intel Xeon
- 4th Generation Sapphire Rapids
- 5th Generation Emerald Rapids
- 6th Generation
  - P – Granite Rapids
  - E – Sierra Forest
- Future Generation
  - P – Diamond Rapids
  - E – Clearwater Forest





Granite Rapids (P Cores)
- Intel 3 node
- Up to 128 Cores / 256 Threads
- 12 x DDR5-6400, 500 W

Sierra Forest (E-Cores)
- Intel 3 node
- Up to 288 Cores
- 12 x DDR5-6400, 400+ W

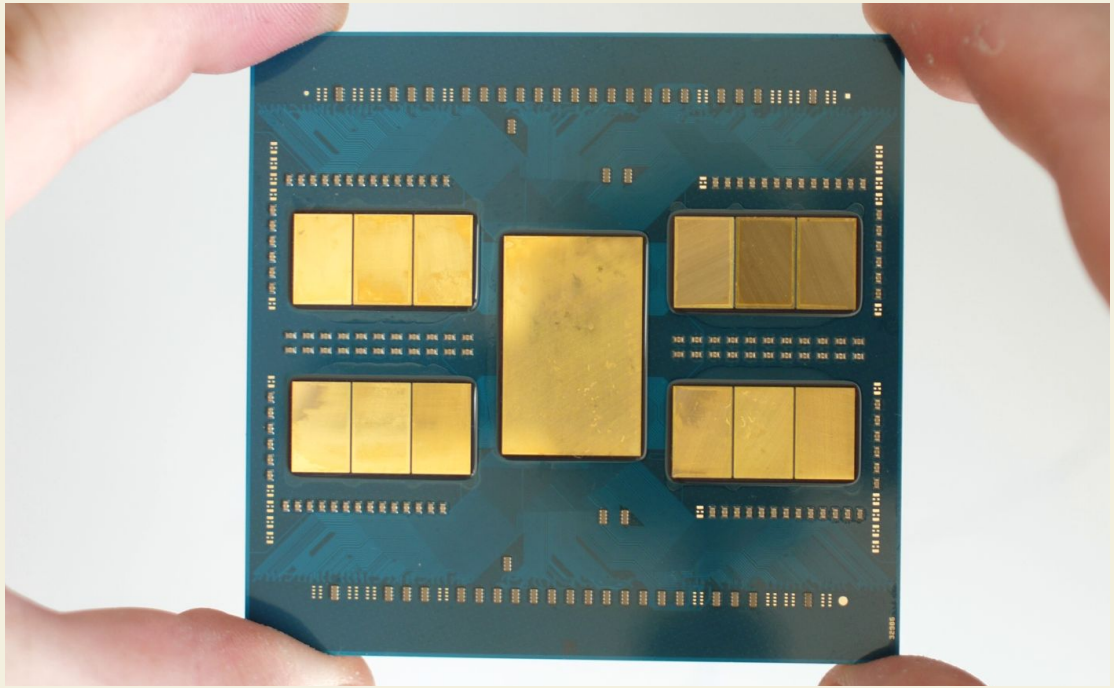


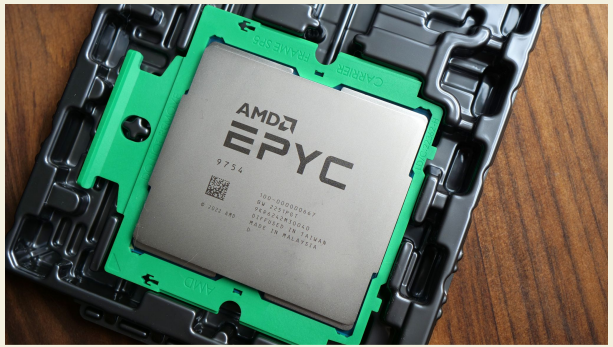
AMD CPU

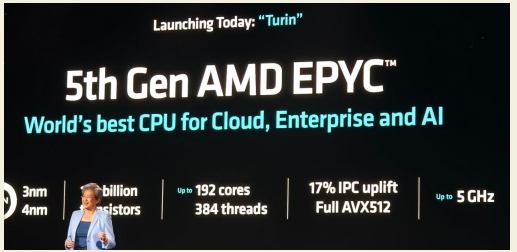
AMD EPYC
- 9001 - Naples
- 9002 - Rome
- 9003 - Milan
- 9004 - Genoa
- 9004 (c) - Bergamo
- 9005 - Turin and Turin Dense

Turin
- TSMC N4X
- Up to 128 cores / 256 threads
- 12 x DDR5-6400, 500 W

Turin Dense
- TSMC N3E
- Up to 192 cores / 384 threads
- 12 x DDR5-6400, 500 W







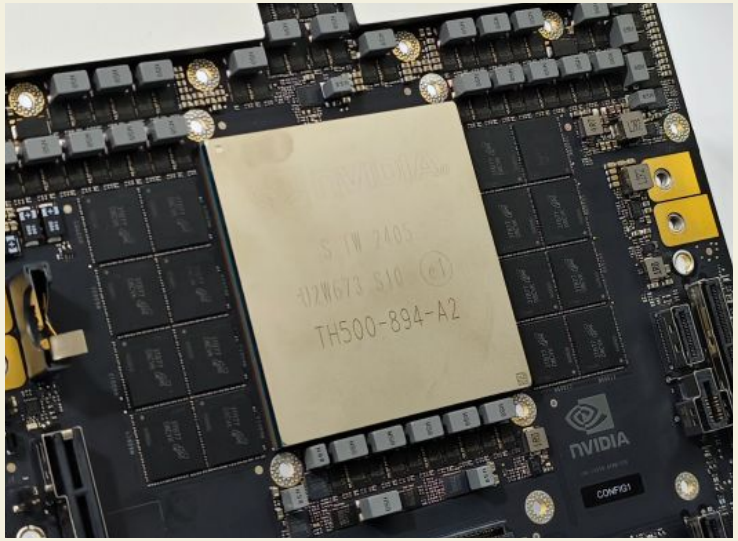
Arm CPU

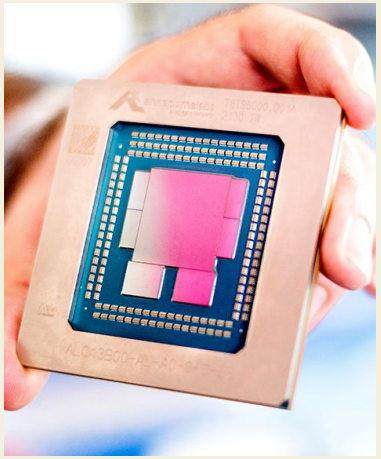
Arm Neoverse N
- N1 - Ares (A76 Derived)
- N2 - Perseus (A710)

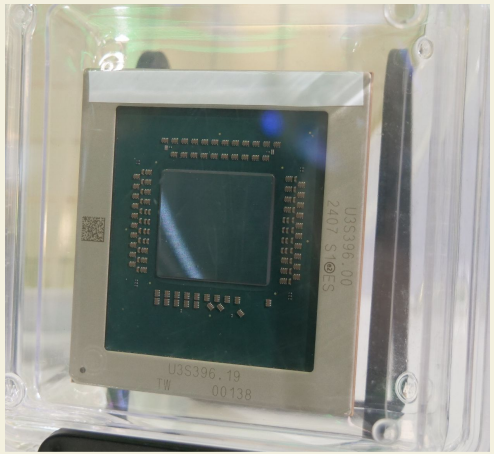
Arm Neoverse V
- V1 - Zeus (X1)
- V2 - Demeter (X3)
- V3 - Poseidon

Arm Neoverse E
- E1 - A65AE derived
- E2 - A510 derived

Popular CPUs using Arm cores
- V2: NVIDIA Grace
- V2: AWS Graviton 4
- V2: Google Axion
- N2: Microsoft Cobolt
- V1: AWS Graviton 3
- N1: AWS Graviton 2
- Custom: AmpereOne

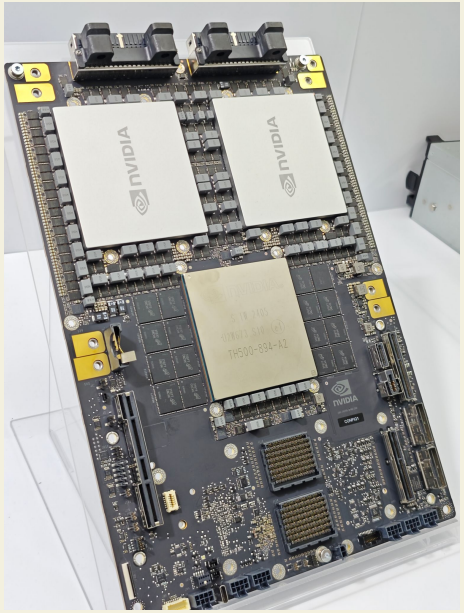






NVIDIA GPU

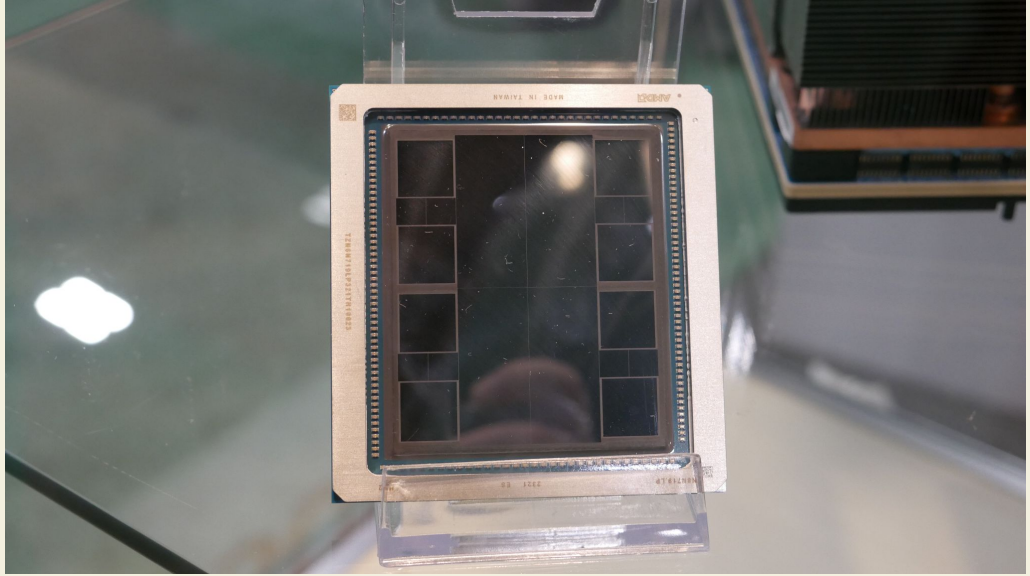
2027:	Rubin Ultra	+ 12x HBM4
2026:	Rubin	+ 8x HBM4
2025:	Blackwell Ultra	+ 8x HBM3e 12H
2024:	Blackwell	+ 8x HBM3e
2023:	Hopper	+ 6x HBM3e
2022:	Hopper	+ 6x HBM3
2020:	Ampere	+ 6x HBM3
2018:	Volta	+ 4x HBM2

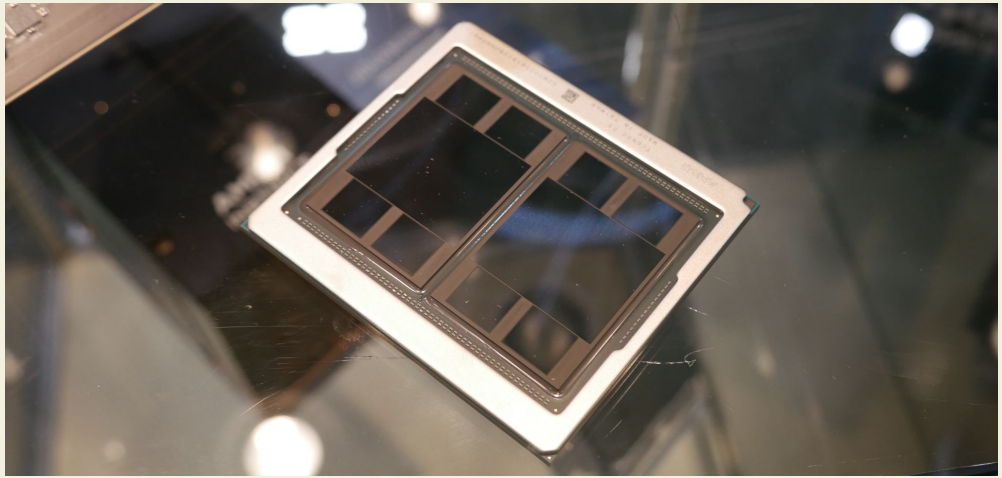


AMD GPU

AMD Instinct

2025:	MI350X	
2024:	MI325X	+ 8x HBM3e 12H
2023:	MI300X	+ 8x HBM3e 8H
2022:	MI250X	+ 8x HBM3
2021:	MI210	+ 4x HBM3
2020:	MI100	+ 4x HBM2
2018:	MI50	+ GDDR6



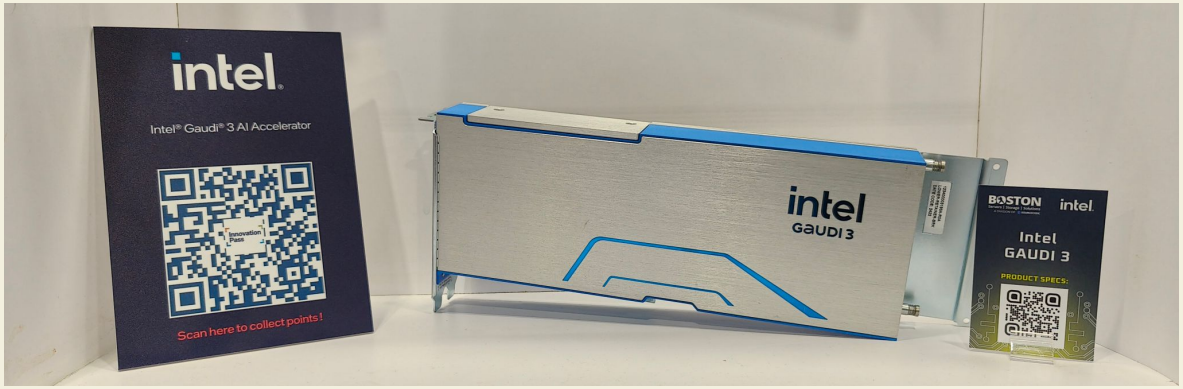


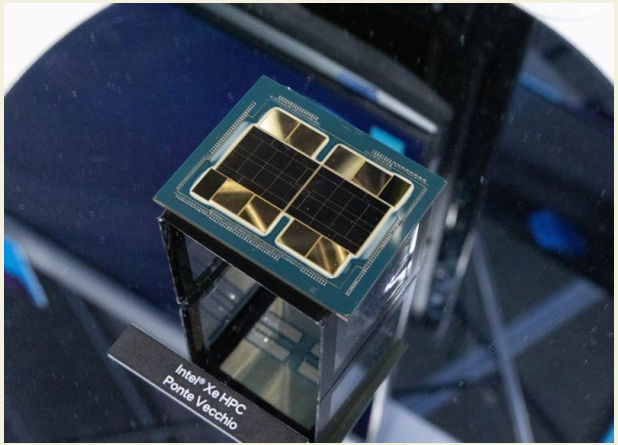
Intel GPU

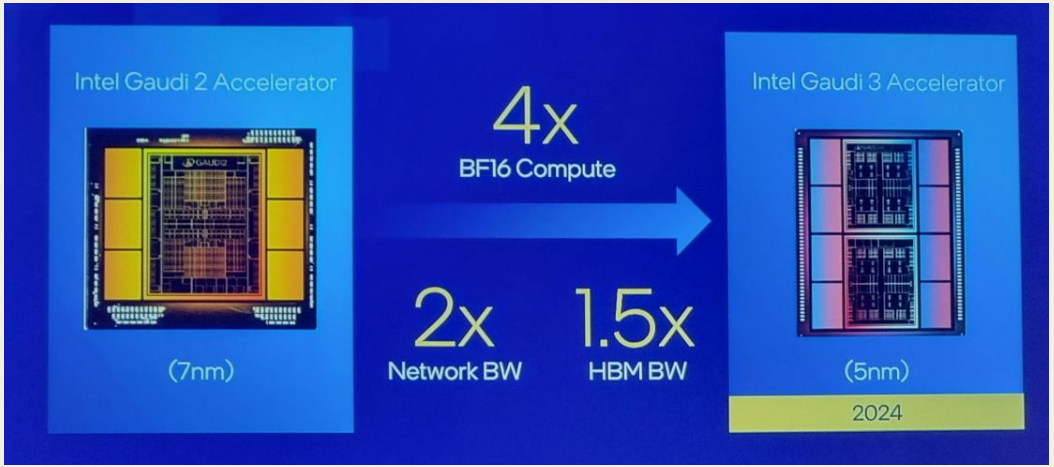
Future: Falcon Shores

HPC: Ponte Vecchio

AI
- 2025: Gaudi 3
- 2022: Gaudi 2

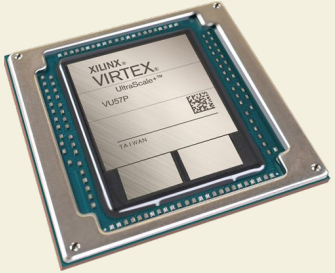


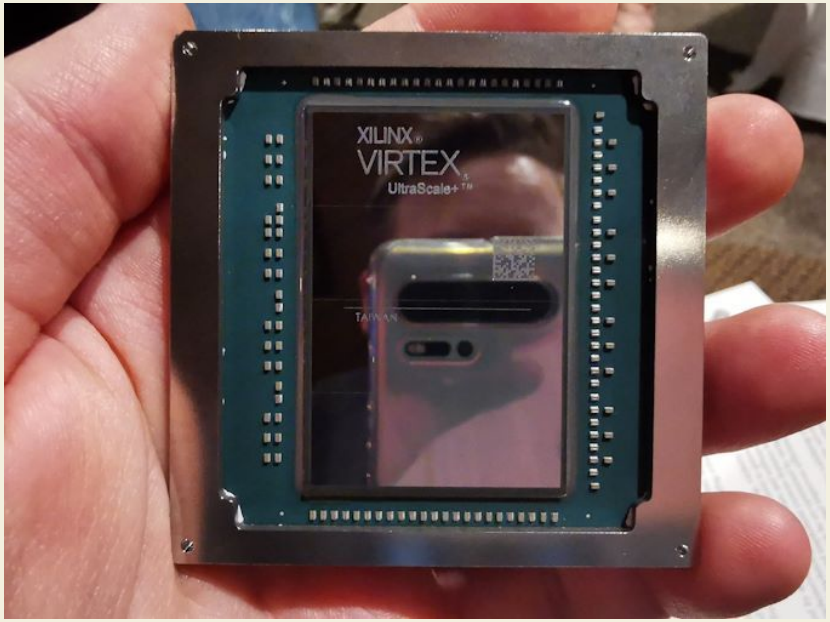


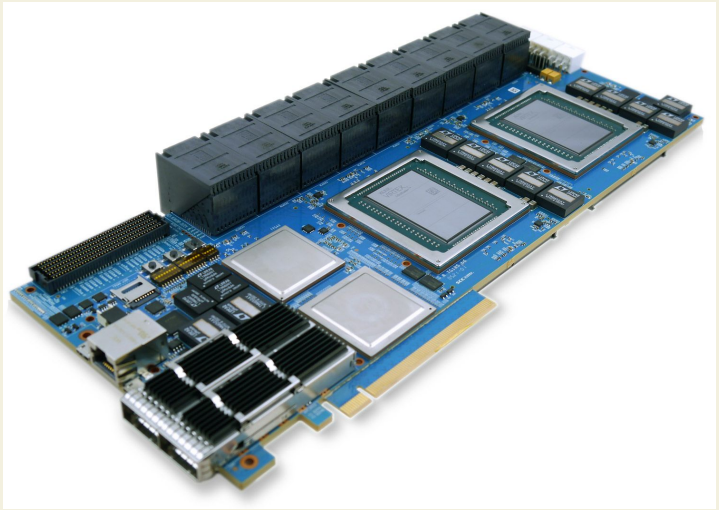


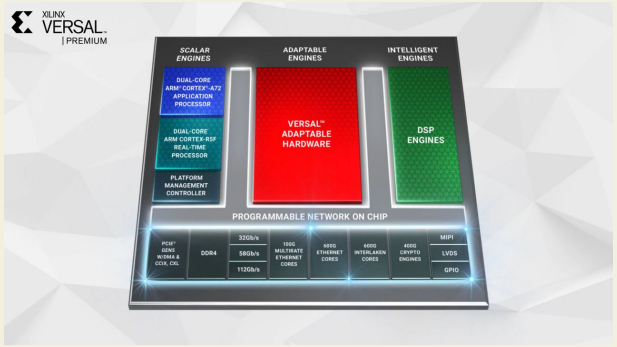
AMD FPGA

Formerly Xilinx
Main brands: Versal, Virtex



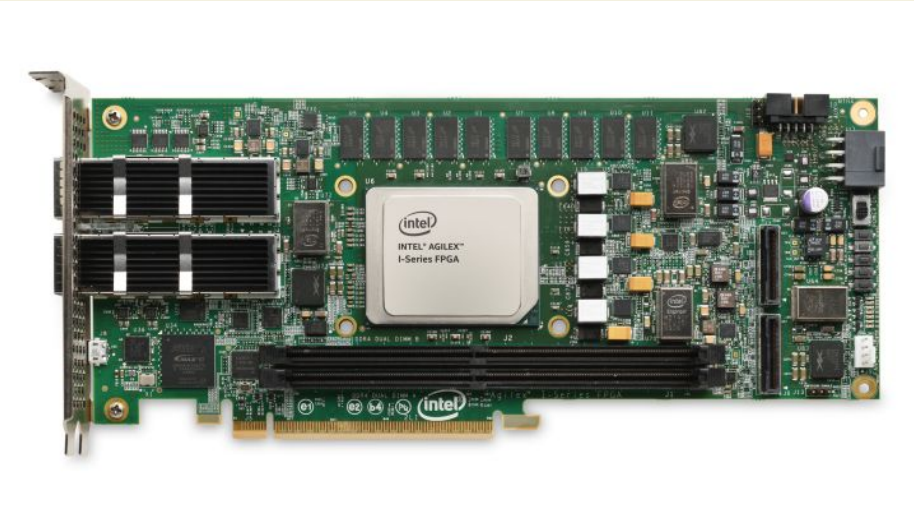


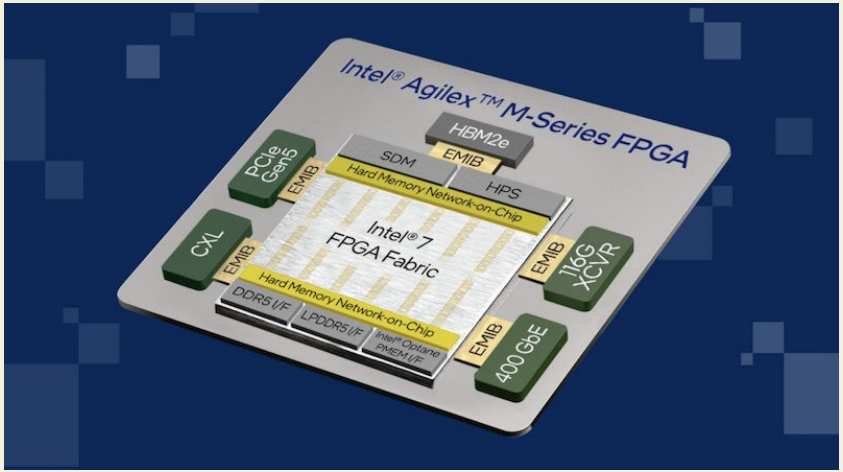


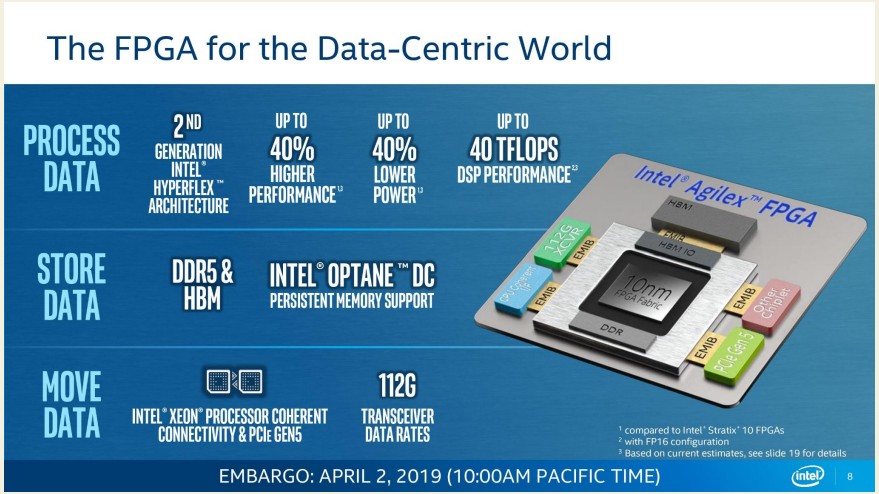


Intel FPGA

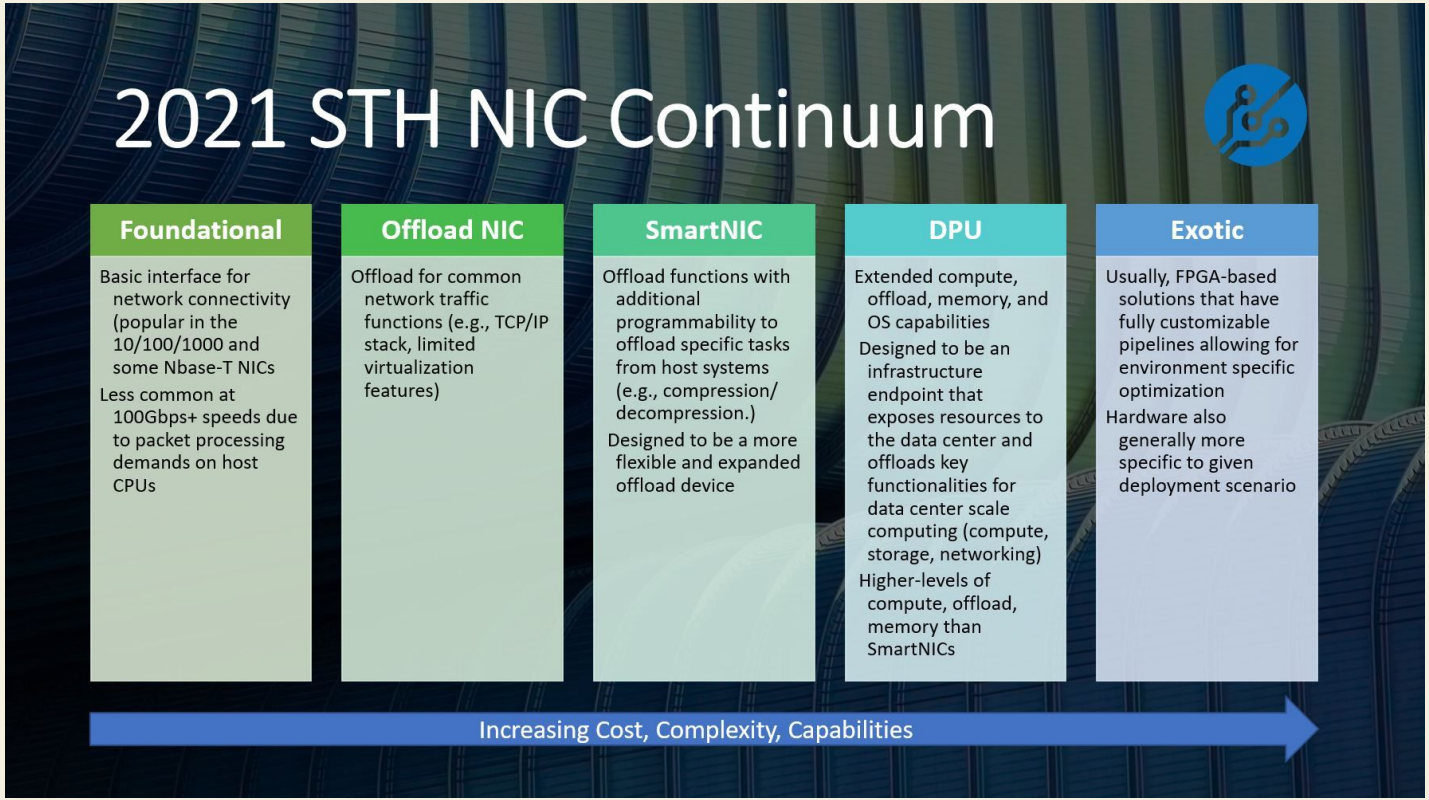
Recently Spun-out again as Altera
Main brand: Agilex





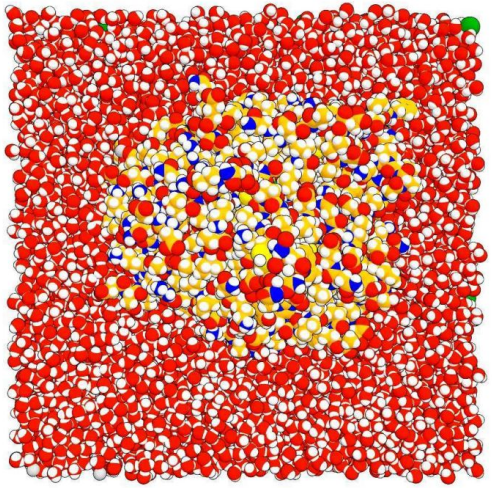


SmartNICs


2021 STH NIC Continuum
Foundational
Basic interface for network connectivity (popular in the 10/100/1000 and some Nbase-T NICs
Less common at 100Gbps+ speeds due to packet processing demands on host CPUs
Offload NIC
Offload for common network traffic functions (e.g., TCP/IP stack, limited virtualization features)
SmartNIC
Offload functions with additional programmability to offload specific tasks from host systems (e.g., compression/ decompression.)
Designed to be a more flexible and expanded offload device
DPU
Extended compute, offload, memory, and OS capabilities
Designed to be an infrastructure endpoint that exposes resources to the data center and offloads key functionalities for data center scale computing (compute, storage, networking)
Higher-levels of compute, offload, memory than SmartNICs
Exotic
Usually, FPGA-based solutions that have fully customizable pipelines allowing for environment specific optimization
Hardware also generally more specific to given deployment scenario
Increasing Cost, Complexity, Capabilities

DEShaw's Anton 3

Molecular dynamics (MD) simulation



- Understand biomolecular systems through their motions
- Numerical integration of Newton's laws of motion
  - Model atoms as point masses
  - Compute forces on every atom based on current positions
  - Update atom velocities and positions in discrete time steps of a few femtoseconds
- Force computation described by a model: the force field

D E Shaw Research

3

A New Chiplet Era

Legacy Hardware: CPU, GPU, FPGA

Old and New: Analog, Neuro, Optical

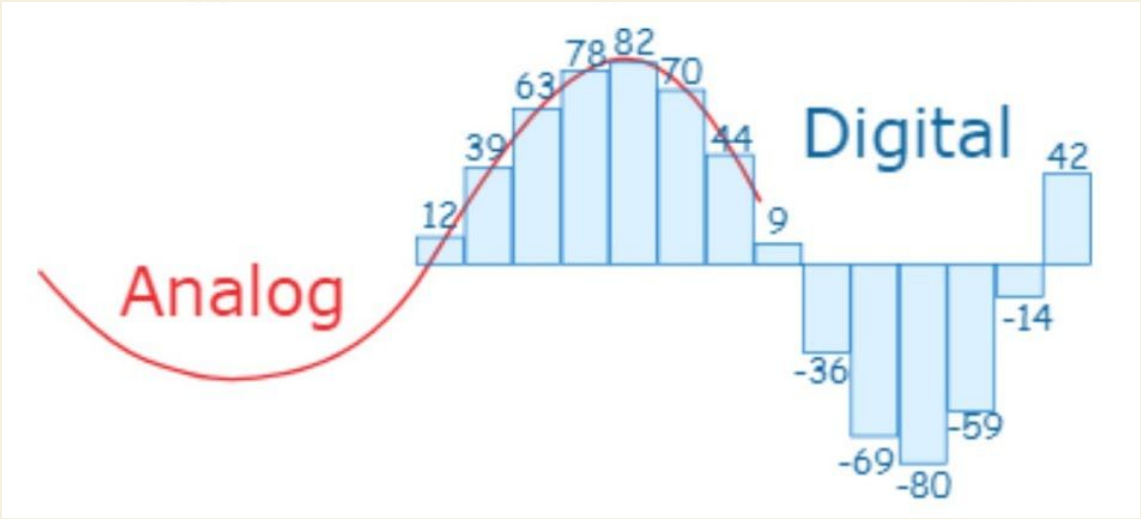
Reduced Precision

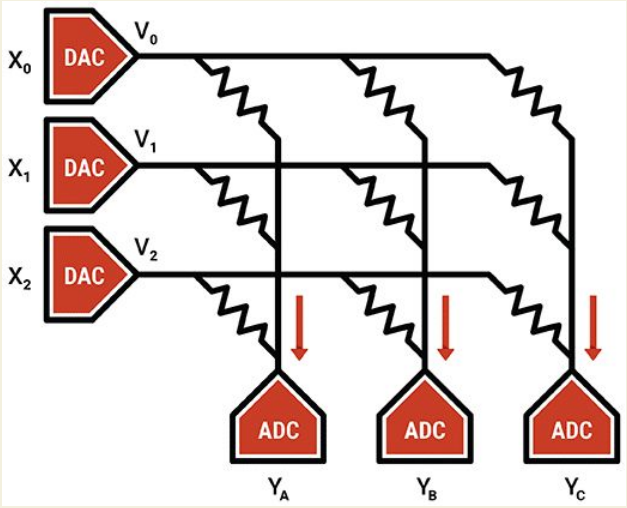
Silicon Market: Datacenter, Edge

AI Hardware: Startup Funding

AI Hardware: What's New

Old and New: Analog Computing





⏫ Super Low Power
⏫ Super Low Latency
⏫ 'Any' Value Possible

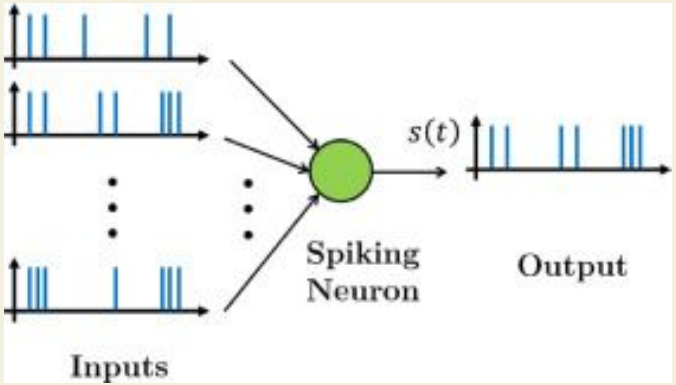
⏬ Conversion Accuracy
⏬ Non-Linear Response
⏬ Scaling

Recent Key Players

- Mythic AI
- IBM
- Aspinity
- EnchargeAI

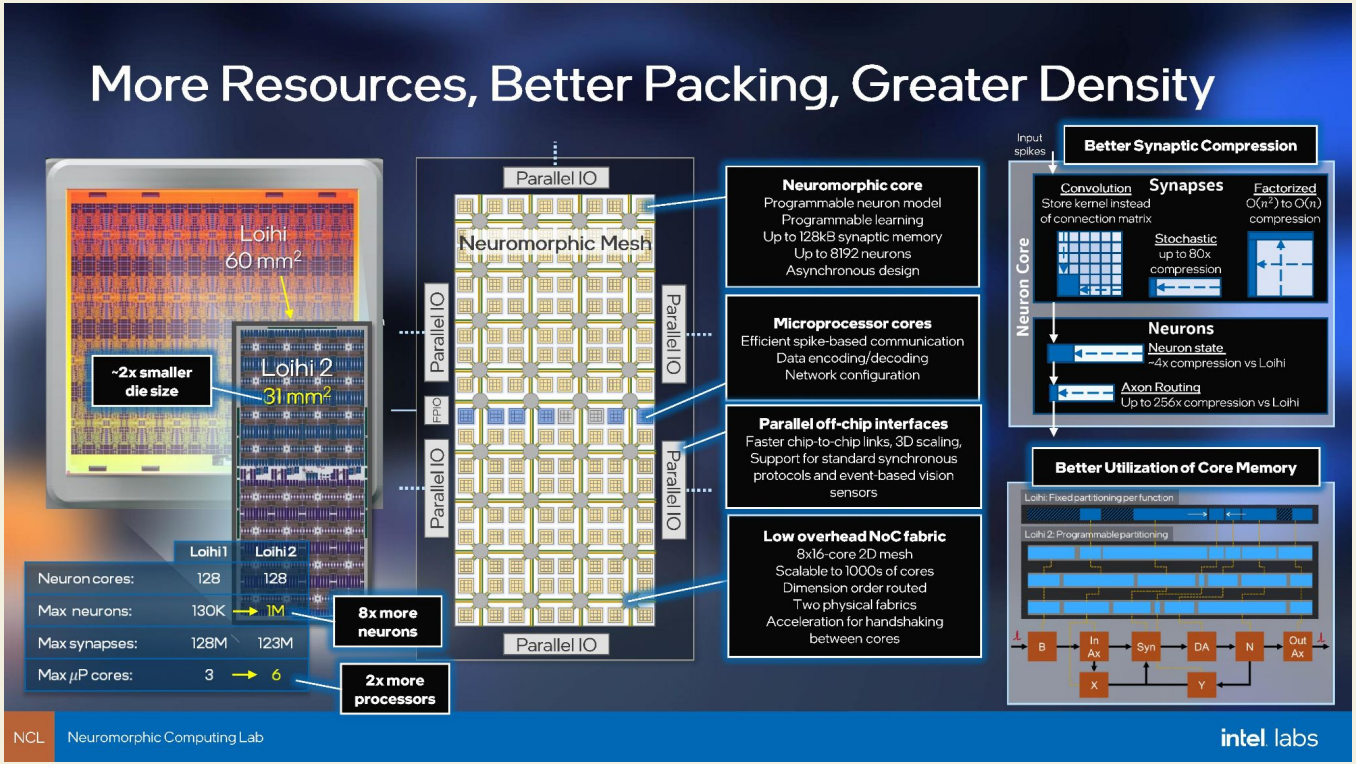
Old and New: Neuromorphic Computing

Beware! Some companies calling themselves ‘Neuromorphic’ aren’t actually doing it

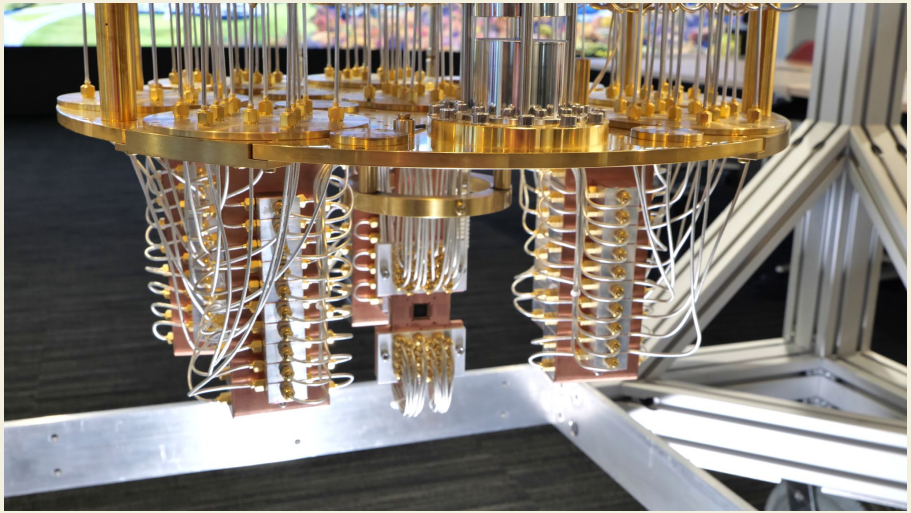


Recent Key Players: Intel Loihi 2, Spinnaker

Intel Loihi 2



Quantum Computing

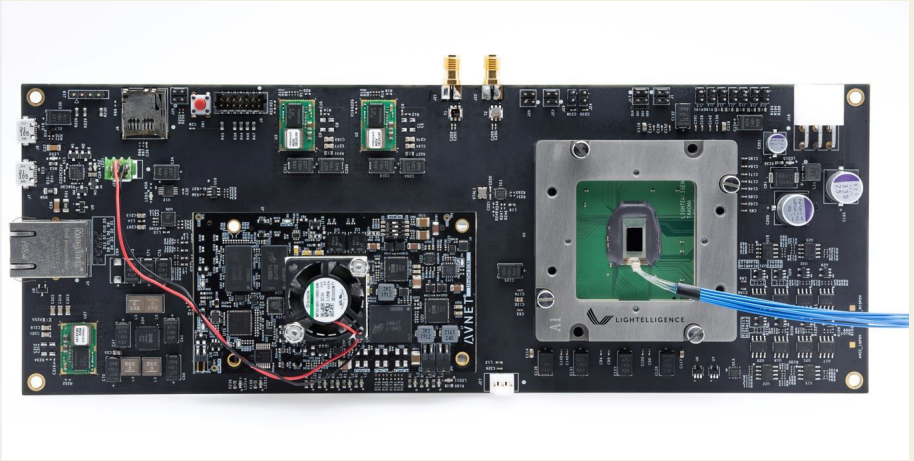


⏫ Physics, Chem, Bio
⏫ Math + Encryption
⏫ Machine Learning

⏬ Any other math
⏬ High Barrier to Entry
⏬ Need a billion qubits

Recent Key Players:
Intel, IBM, Google,
Microsoft, Amazon,
DWave, Alibaba, IonQ

Optical Computing



⏫ No Power
⏫ Speed-of-light fast

⏬ Scales Poorly
⏬ Manufacturing

Recent Key Players:
LightMatter, Lightelligence

A New Chiplet Era

Legacy Hardware: CPU, GPU, FPGA

Old and New: Analog, Neuro, Optical

Reduced Precision

Silicon Market: Datacenter, Edge

AI Hardware: Startup Funding

AI Hardware: What's New

Quantization

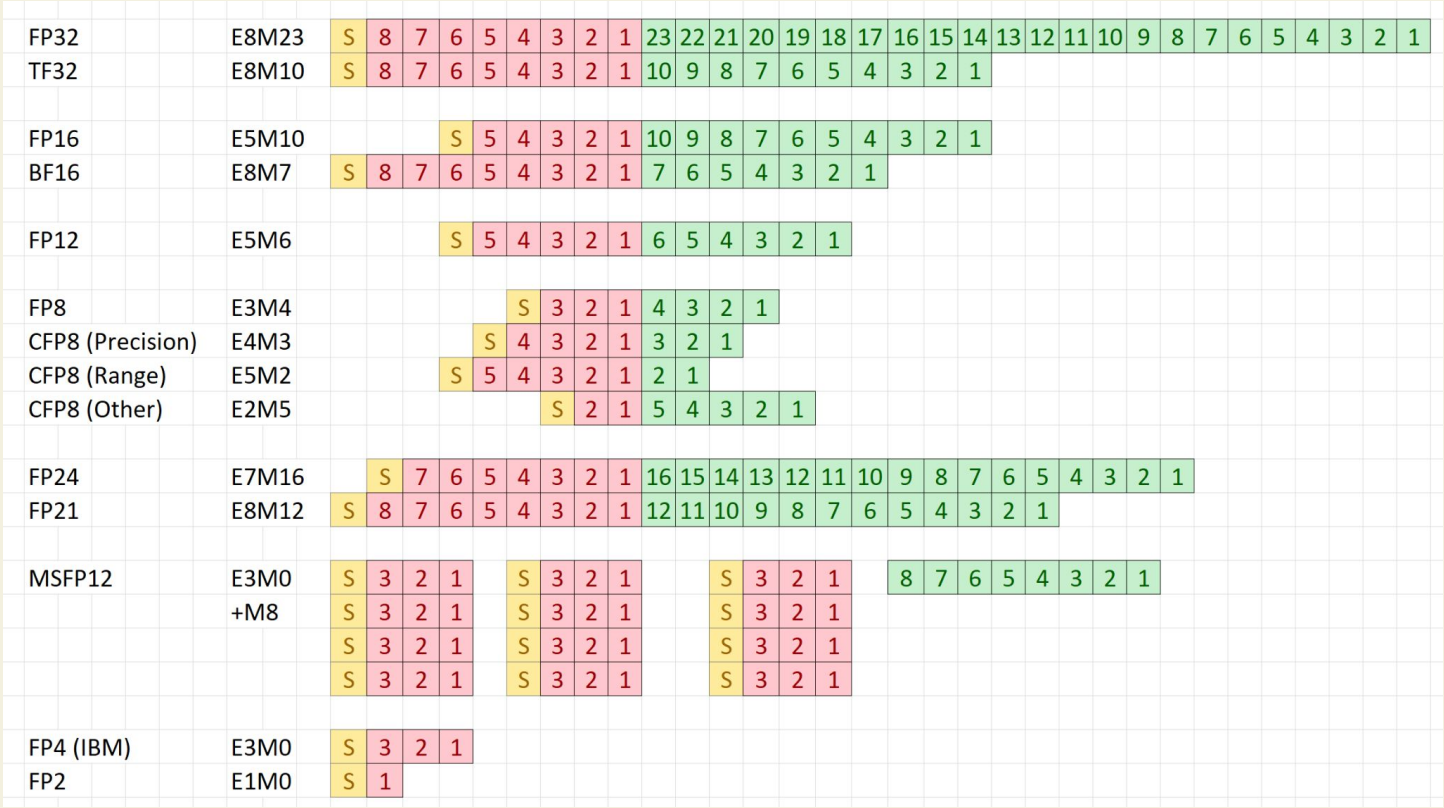
Using fewer bits saves power, and with the right architecture, can be sped up. But the tradeoff is range and accuracy.

Quantization

Power savings are realised:

Data Type	Bits	FLOPS/Joule
FP8	8	33.0x
BF8	8	33.0x
BF16	16	13.0x
FP16	16	13.0x
TF32	32	1.8x
FP32	32	1.8x
FP64	64	1.0x

Quantization



A New Chiplet Era

Legacy Hardware: CPU, GPU, FPGA

Old and New: Analog, Neuro, Optical

Reduced Precision

Silicon Market: Datacenter, Edge

AI Hardware: Startup Funding

AI Hardware: What's New

Silicon Market: Four Battlegrounds

Datacenter Training (DCT)

Giant clusters of large (10k+) chips working together

Consume Megawatts

Develop Foundation Models

Datacenter Inference (DCI)

Giant clusters of PCIe devices working independently

Dealing with lumpy workloads

Traditional Cloud business model

Localised Inference (LI)

On-premise hardware, often PCIe

Dealing with in-network requests during office hours

Sized down version of DCI

Edge Inference (EI)

Small custom ASICs not on PCIe (eg Wearables, XR, IoT)

Running small models, non-LLM

Often for specific features

Silicon Market: Four Battlegrounds

Datacenter Training (DCT)

NVIDIA H100, H200, GB200

AMD MI350X, MI300X, MI250X

Sambanova SN40L, Cerebras WSE3

Datacenter Inference (DCI)

NVIDIA L40	Tenstorrent	Rebellions
AMD MI210	Groq LPU	FuriosaAI
Intel Gaudi 3	IBM AIU	Neuchips
Google TPU	Untether	d-Matrix

Localised Inference (LI)

Qualcomm Cloud AI 100

DeepX

Flex Logix

Edge Inference (EI)

Recogni	Sima.ai	Synsense
Axelera	Syntiant	Innatera
Kneron	XMOS	GrAI
Hailo	Encharge	EdgeQ

A New Chiplet Era

Legacy Hardware: CPU, GPU, FPGA

Old and New: Analog, Neuro, Optical

Reduced Precision

Silicon Market: Datacenter, Edge

AI Hardware: Startup Funding

AI Hardware: What's New

Startup Funding

150+ AI Hardware Startups.

Main Questions:

- **Consolidation when?**
- **IPO when?**

Startup Funding

In 2024:

- One IPO
- Two Acquisitions
- One Merger
- Two Deaths

But:

- $3.00 Billion Funding Raised
- 10 companies $100m+
- I track $17B+ in this market

Startup Funding

In 2024:

- One IPO
- Two Acquisitions
- One Merger
- Two Deaths

But:

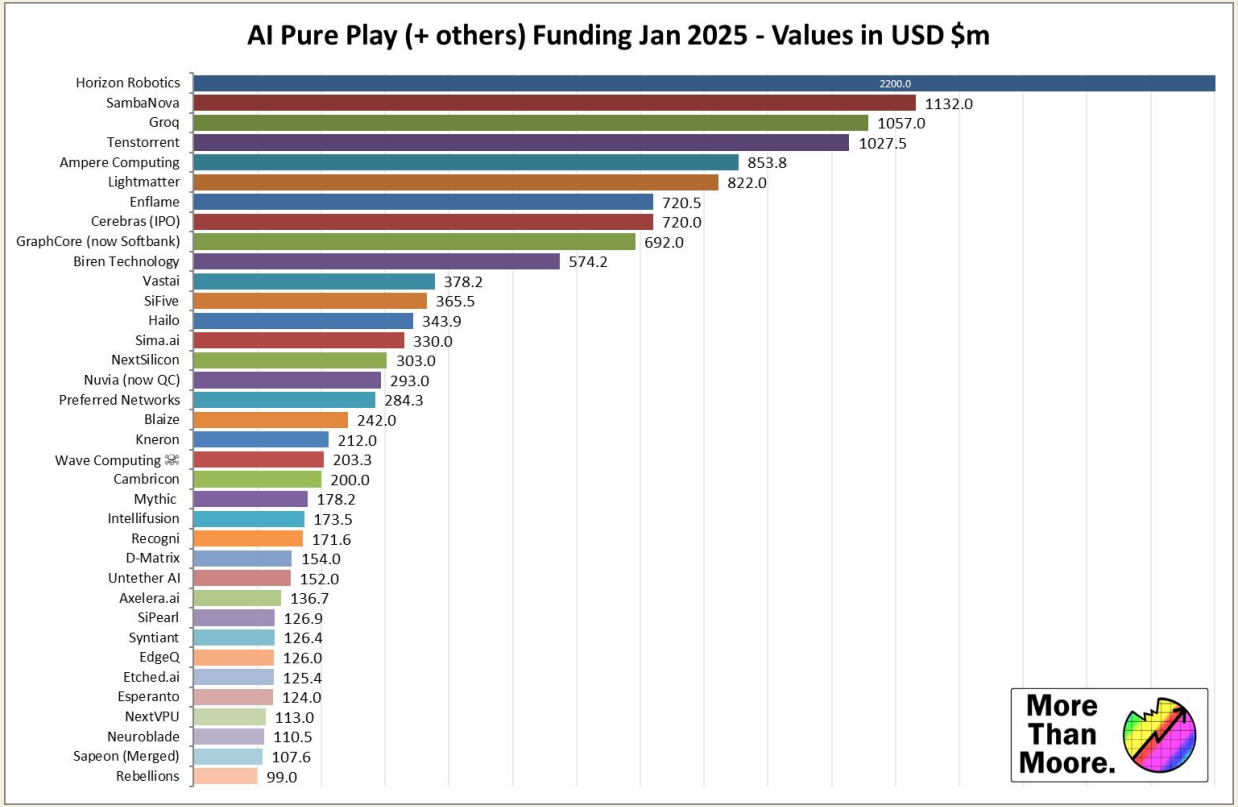
- $3.00 Billion Funding Raised
- 10 companies $100m+

21 companies had public funding rounds

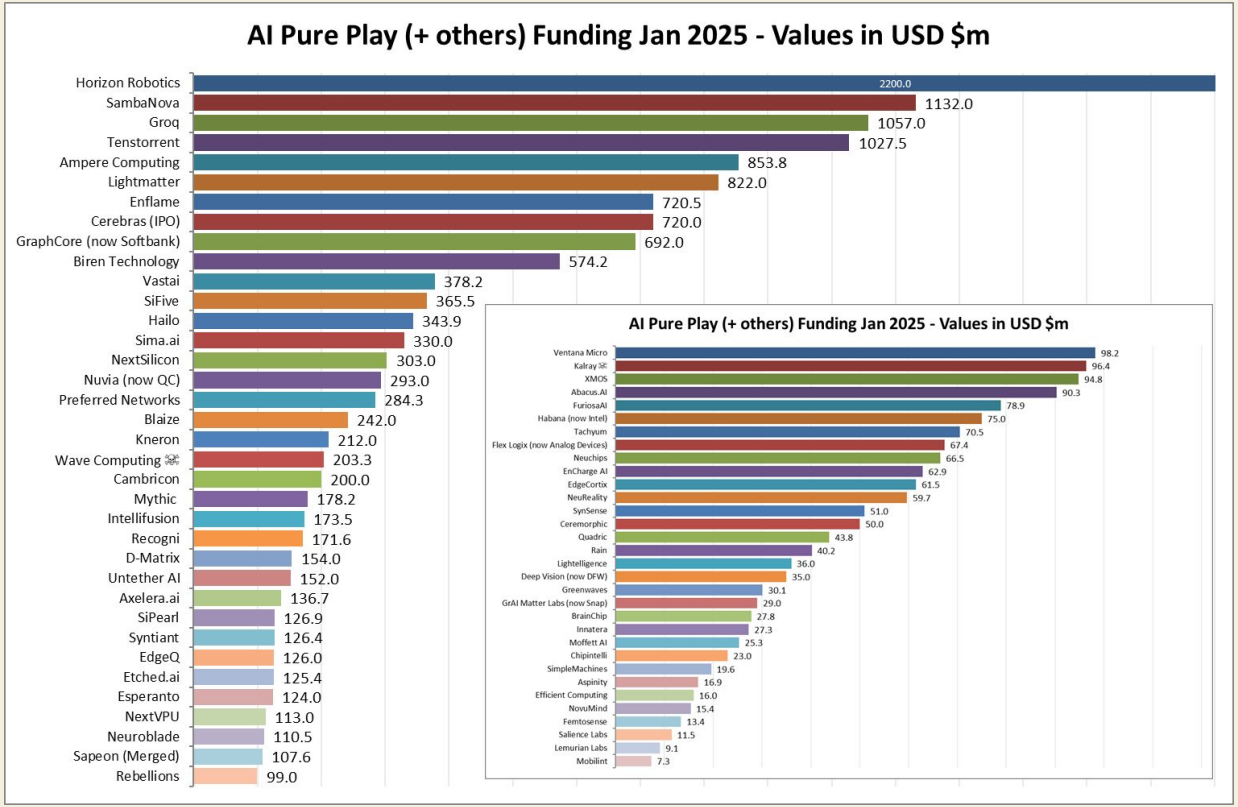
Top Ten in 2024

1.	Tenstorrent	$693m
2.	Groq	$690m
3.	Lightmatter	$400m
4.	Ayar Labs	$155m
5.	Sima.ai	$140m
6.	Preferred Networks	$138m
7.	Etched.ai	$120m
8.	Hailo	$120m
9.	NextSilicon	$100m
10.	Recogni	$100m

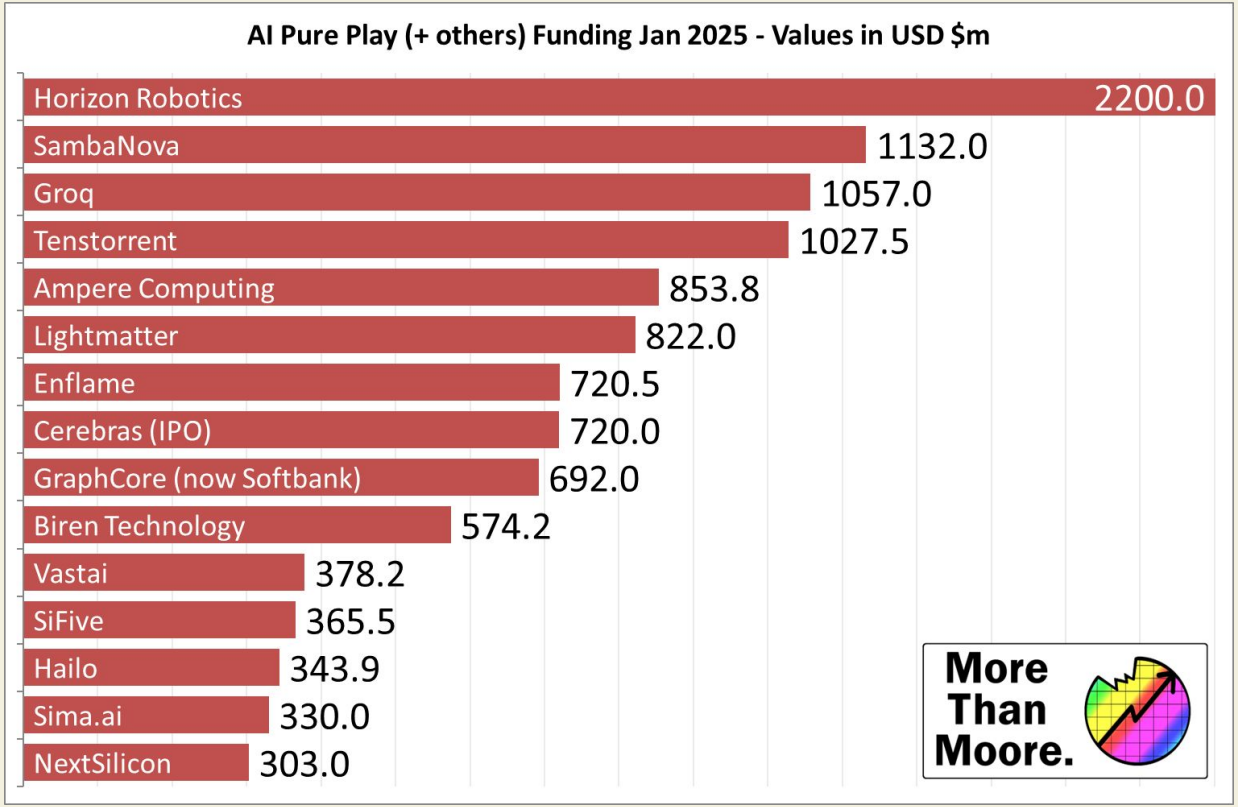
Startup Funding



Startup Funding



Startup Funding



Who's Hot: Tenstorrent

Founded in 2015
Raised $1.03 billion
Current CEO is Jim Keller, ex-AMD, Tesla, Intel, Apple

Portfolio has two main product families:

- High-performance RISC-V Core (Ascalon)
- High-performance AI Core (Tensix)

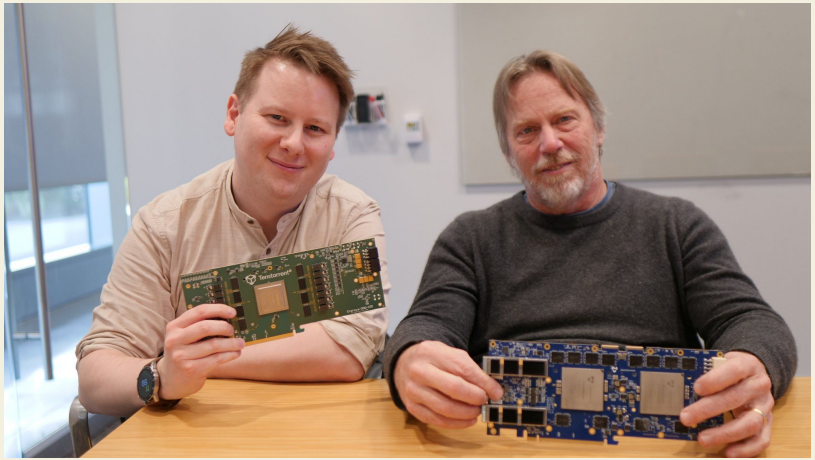
Both are offered in two different ways

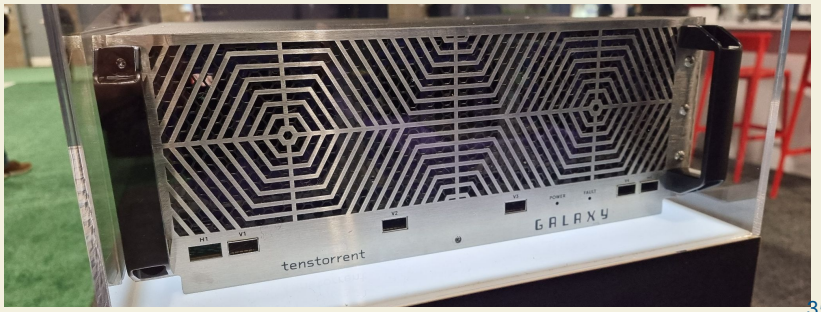
- Licensable IP
- Chiplets or full product: PCIe to Server

Major licensing deals already struck with LG, Hyundai, Samsung.
Working very closely with Rapidus, Japanese 2nm Fab efforts

Anyone can go buy Tenstorrent's AI PCIe card. Open source software, fully dynamic community, aiming for software standards.

Galaxy servers with 32 PCIe cards are already being installed.





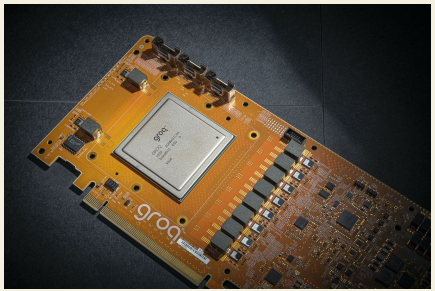
Who's Hot: Groq

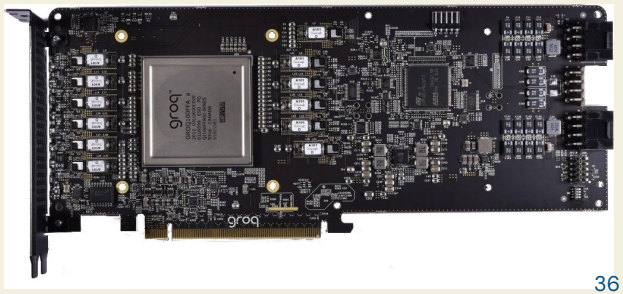
Founded in 2016
Raised $1.06 billion
Current CEO is Jonathan Ross, ex-Google

Main Product is Groq LPU, a 2021 data-flow chip
Business model used to be hardware, now is mostly a cloud
Regularly competes for fastest datacenter inference LLM speed, but requires lots of power.

Will release new chip in 2025 built on Samsung LP4X
Expected to be a larger design but with stacked SRAM or DRAM
Solves critical issue of first chip - not enough onboard memory

Very enthusiastic marketing team!





Who's Hot: Cerebras

Founded in 2016
Raised $720 million - has filed for IPO in Sep 2024
Current CEO is Andrew Feldman

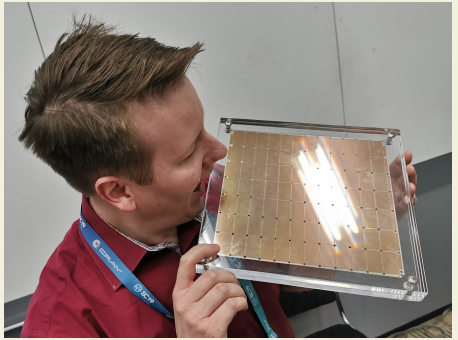
Saw that a critical aspect to ML performance scaling was chip-to-chip bandwidth, so designed a chip as big as a wafer. The wafer scale engine.

A single WSE is a full wafer - redundancy provides near 100% yield.
Custom solution with custom power and cooling - fits in a 19-inch rack.

Newest gen WSE3 has:

- 900,000 cores
- 1.2 Tbps off-chip bandwidth
- 44 GB on-chip SRAM
- Built on TSMC N5
- 21 PetaBytes/sec memory bandwidth (vs 0.003 on H100)

Made ~$1billion revenue from a single customer who bought 576 machines across 9 datacenters.



What Happened to: Graphcore

Founded in 2016
Raised $692 million - acquired by Softbank in 2024
Still operates as a separate group

Built several chips (IPUs) to tackle computer vision-style workloads
Latest generation chip was unable to scale to Transformers
Wasn't competitive against other chips for Large Language Models

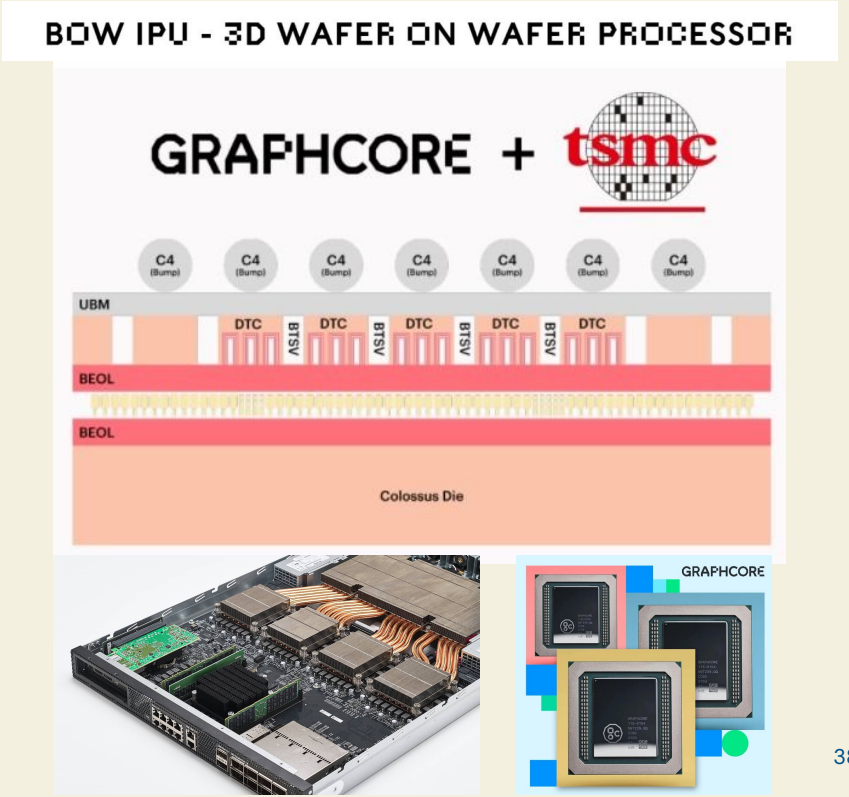
Had an impressive roadmap

- 2020: TSMC 7nm chip
- 2022: Lead TSMC partner for custom packaging of 2020 chip
- 2024: TSMC 3nm chip*

Unfortunately, the 2024 chip never materialized - no large 3nm chips have been made yet. Little interest in 2022 update, not competitive for LLMs.

Lost large deals with Microsoft, Dell.

Ran out of money. Acquired by Softbank for approx investment.
Reminder: Softbank also has majority ownership of Arm



What Happened to: Mythic AI

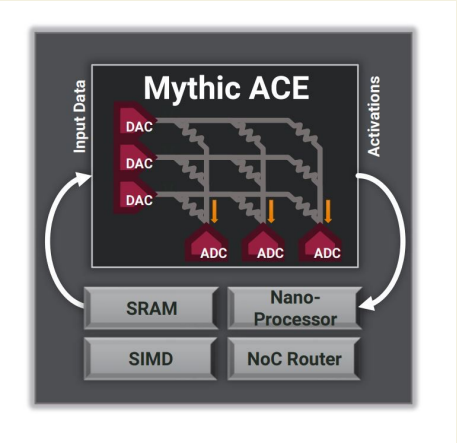
Founded in 2012
Raised $178 million
‘Died’ 2022 due to lack of money, small $10m injection in 2023

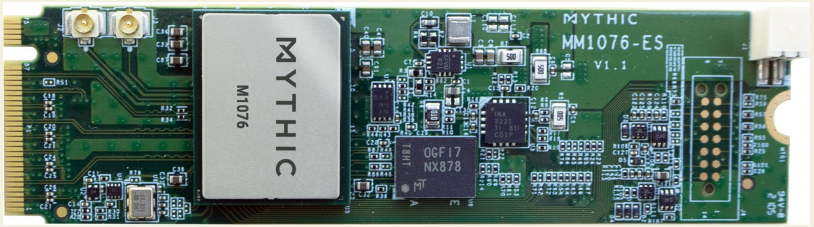
Focused on analog compute-in-memory solutions for embedded up to data-center inference. Due to early cycle, focused on computer vision and non-transformer models. Interest from Lockheed Martin and other defence contractors.

Why did Mythic AI fail? It ran out of money.

Reports suggest the company tried to do too much itself, while competing in a competitive field, and simply overspent before achieving proper sustainable revenue.

After the $10m injection in 2023, the company appointed a new CEO in August 2024. Latest reports show a defence-only focus, and no-outreach.





A New Chiplet Era

Legacy Hardware: CPU, GPU, FPGA

Old and New: Analog, Neuro, Optical

Reduced Precision

Silicon Market: Datacenter, Edge

AI Hardware: Startup Funding

AI Hardware: What's New

Three Topics Dominating Conferences

Advanced Packaging

- Co-Packaged Chiplets (2D)
- Large Silicon Interposers (2.5D)
- Small Silicon Bridges
- Large Organic Interposers
- Logic-on-SRAM Stacking (3D)
- SRAM-on-Logic Stacking
- Logic-on-Logic Stacking (Future)
- Combination of above (3.5D)
- Stacked HBM on Logic (HBM4)
- Co-packaged Optics (Future)
- Optical Waveguide Interposers (Future)
- Heterogeneous Integration (sub-10 micron)
- Glass Substrates (Future)

TSMC
- CoWoS-S, CoWoS-L, SoIC-X

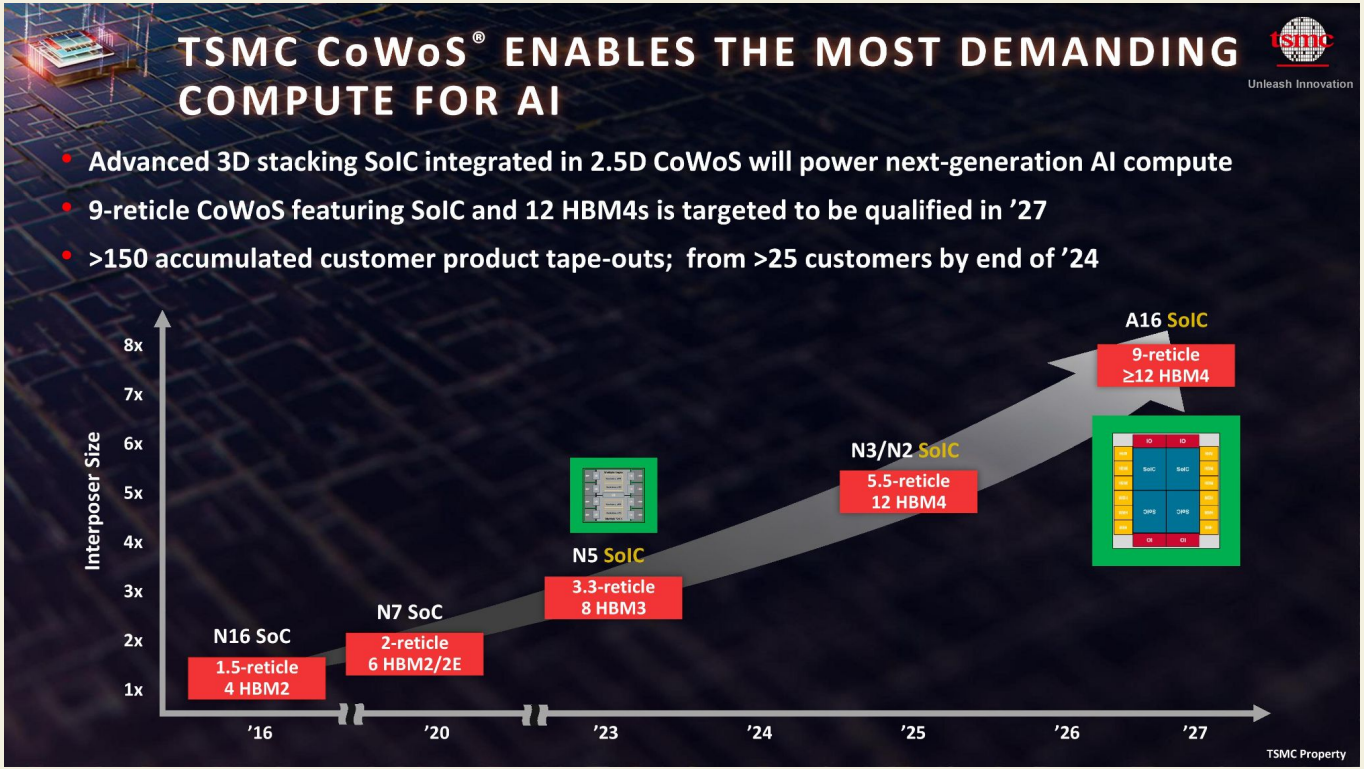
Intel
- EMIB, Foveros Direct, Foveros Omni

Samsung
- I-Cube, X-Cube, H-Cube

Also OSATS and Supply Chain Enablement
- ASE, Amkor, GlobalFoundries, JCET
- Applied Materials, LAM Research, TEL, Xperi

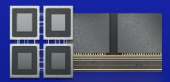
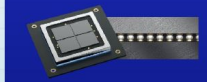
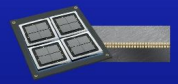
Research
- Imec, Fraunhofer, Leti, IBM

Advanced Packaging



Advanced Packaging

Continuing the Packaging Evolution at Intel

2D/MCP	2.5D	2.5D/3D	Hybrid Bonding	Next Gen Interconnects
FCBGA/FCLGA	Embedded Multi-die Interconnect (EMIB)	Foveros	Foveros Direct	Glass Core Substrate
				
bump pitch ~100µm	bump pitch 55→45 →45/36µm w/TSV	bump pitch 36→25µm	bump pitch < 10µm	
▪ Support to 92x92mm size → >100x100mm ▪ Enabled STIM w/BGA	▪ Products shipping since 2017 ▪ High bandwidth / HBM ▪ Supports large form factor: 4.5x reticle equiv. → 6x+ reticle equiv.	▪ First product in 2019, HVM ramp from 2023 ▪ Wafer-level packaging capabilities ▪ Optimized for cost/performance	▪ Direct copper-to-copper bonding for high-density and low resistance interconnects ▪ Goal: Best power per bit performance	▪ Continued feature scaling ▪ Improved power delivery ▪ Enables 448G/ HSIO

Advanced Packaging Era

*All packaging plans and development roadmaps subject to change

Three Topics Dominating Conferences

Connectivity

Between GPUs

- NVLink (NVIDIA)
- UALink (Mostly Everyone Else)
- Ethernet (Startups like Tenstorrent)
- Custom (Google)

Between Servers

- Ethernet (Ultra Ethernet Consortium)
- Infiniband (NVIDIA / Mellanox)
- Omnipath (Cornelis Networks)
- Custom (Amazon, Google)

The importance of high-speed SERDES IP on leading edge process node technologies

Three Topics Dominating Conferences

Co-packaged Optics

Falls under different names

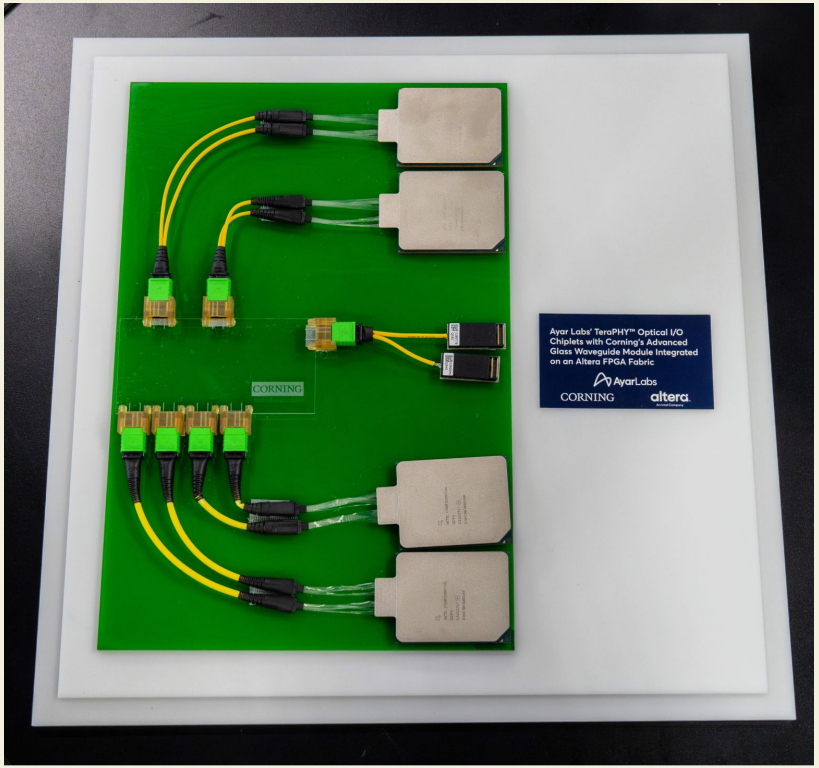
- Silicon Photonics (general)
- Co-Packaged Optics (Broadcom)
- Optical IO (Ayar Labs)

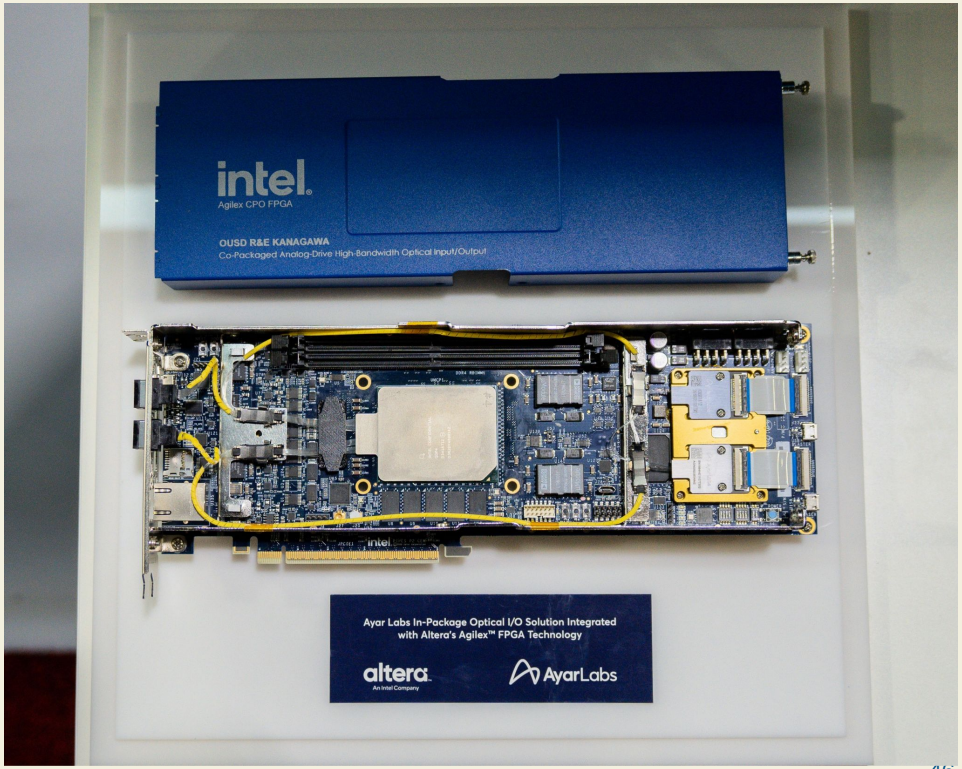
Modern electrical SERDES is limiting - eight way GPU clusters in ML is the norm.
—or
NVIDIA uses custom electrical switching to bring together 72+ GPUs to act like one machine

The solution is direct chip-to-chip optics!

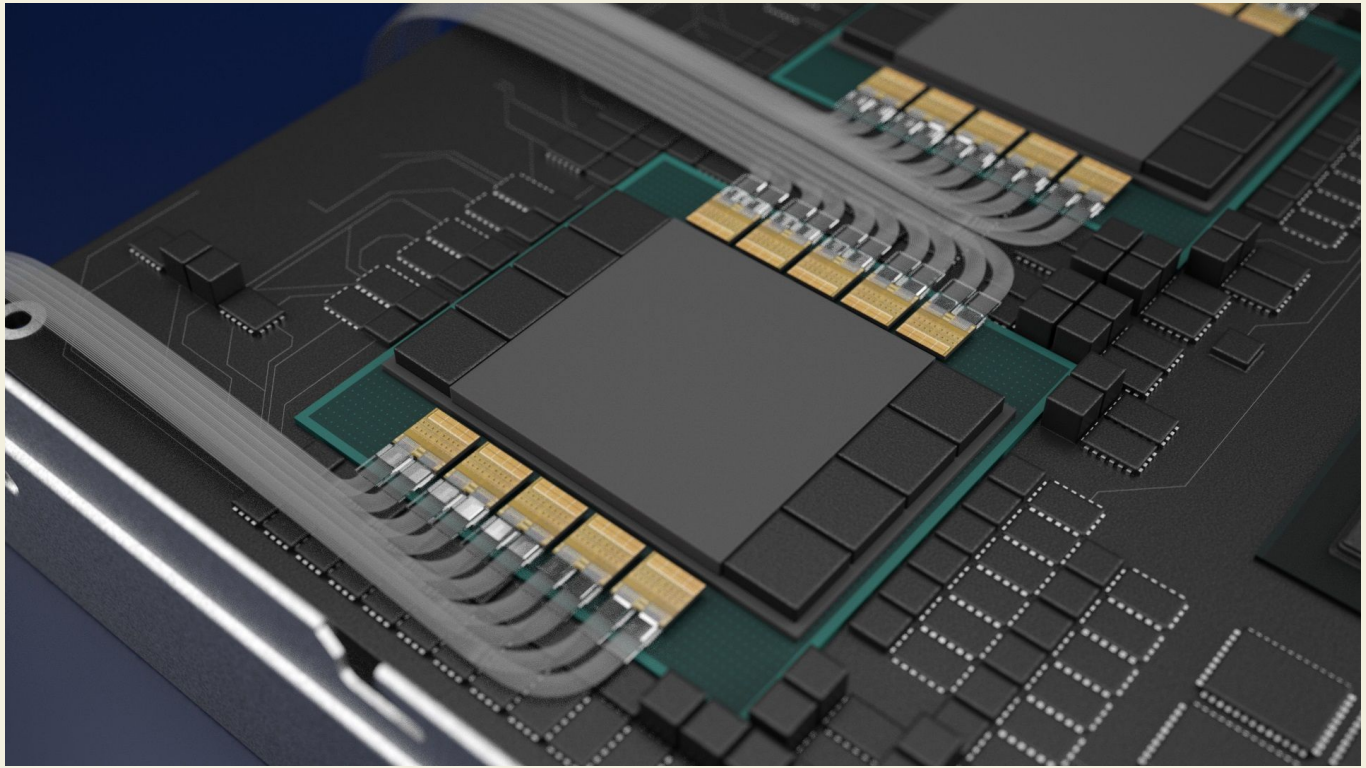
Build an optical chiplet into the ML

Co-Packaged Optics





Co-Packaged Optics

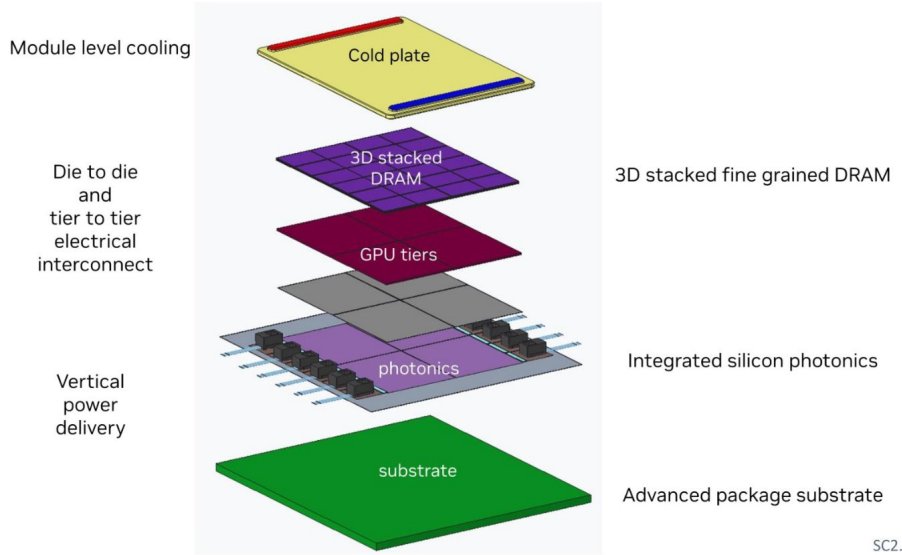


Ayar Labs
SC24

Co-Packaged Optics

Vision for Future AI Accelerators

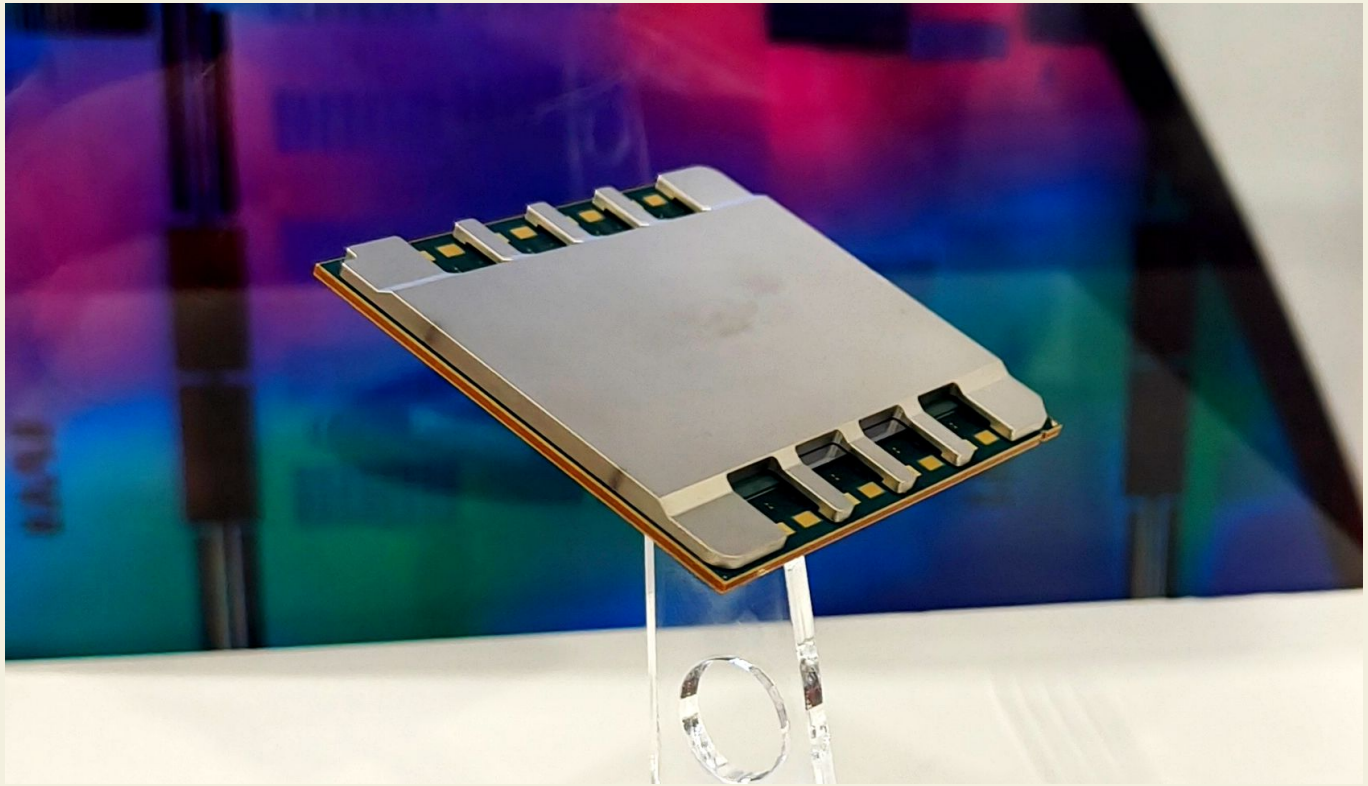
Innovations needed in 2.5D/3D packaging, integrated photonics, power delivery, and thermals



SC2.2 48

NVIDIA
IEDM 2024

Co-Packaged Optics



Lightmatter
SC24

Co-Packaged Optics

Everyone is investing on any solution

- Ayar Labs and Lightmatter raised $555m in 2024

Ayar Series D investors include

- AMD
- NVIDIA
- Intel
- Defence Sector
- Global Foundries

Lightmatter Series D investors include

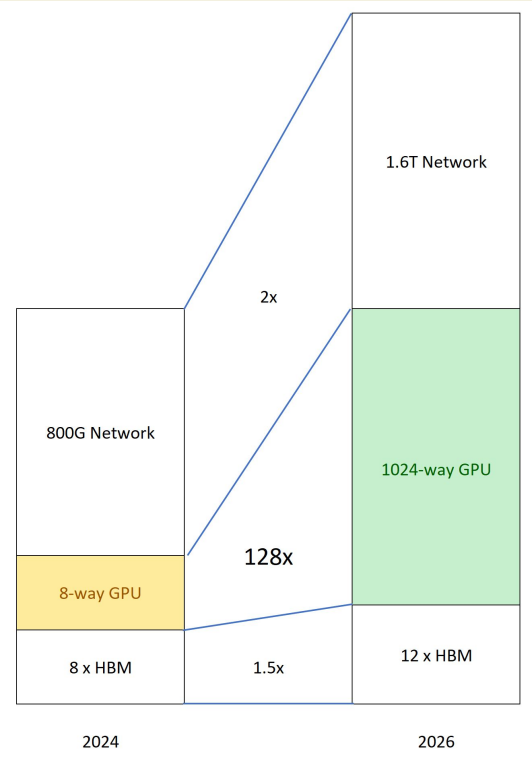
- Google Ventures
- Fidelity
- T.Rowe Price

Expected initial deployment in the 2026-2029 timeframe.

Issues to be solved:

- Co-design with ML ASIC
- Repeatable predictable fiber attach
- Thermal variation
- Manufacturing variation
- Manufacturing volume
- Software stack adjustment

Co-Packaged Optics



By 2026, we expect

- HBM per ASIC to grow from 8 to 12
- 400/800G networks to update to 1.6T

But, the limiting factor is all-to-all ASIC connectivity.

Optical promises to increase the 8-way to 128-way and above, therefore reducing chip-to-chip latency and increasing bandwidth without requiring the 'external network' to waste time and power.

Today's GPUs in training rarely go beyond 35-40% utilization, due to syncronization and data bandwidth limits. Optics aims to solve this.

Side note: Optics can also be used for higher HBM attach in some models.

Next-Generation AI Silicon

Dr. Ian Cutress
More Than Moore